ViTopic: Topic Modeling on Images by Clustering of Vision Transformer Embeddings

Fausto J. German Jimenez
ITCS 5156 – Applied Machine Learning

Computer Science Department
at
The University of North Carolina at Charlotte

Charlotte, North Carolina, USA

May 3, 2023

Abstract

This project proposes ViTopic, a system for generating visual topic models using a combination of vision transformer embeddings, image captioning, and neural topic modeling. ViTopic aims to reduce the time and labor required to analyze image data by automatically clustering images into groups and generating concept-guided descriptions for each collection. The system uses visual embeddings to generate a pair-wise similarity matrix for visual similarity search. It also combines image captions and visual embeddings to generate topic descriptors using neural topic modeling techniques. The results outlined in this report demonstrate ViTopic's effectiveness in generating high-quality image-based topic models. Leveraging ViTopic could allow users to create highly interpretable visual topic models for various applications, such as image search, image classification, and visual sensemaking, without manual labeling.

Keywords: Computer vision, Visual Similarity Search, Recommender Systems, Image Captioning, Unsupervised Learning

Table of Contents

Chapter 1

1 Introduction

Image data is a critical component in many fields, ranging from social media analysis and medical research to environmental studies, agriculture, and the entertainment industry. Deep learning models, such as convolutional neural networks (CNNs) and vision transformers, have significantly improved our ability to analyze and extract insights from image data. However, collecting, sorting, clustering, and labeling vast amounts of images remains a challenging and time-consuming task that requires significant human resources. Although computer vision researchers have recently developed faster and more efficient algorithms for image clustering [2,16], vision-language topic descriptions [7], and even visual summarization of groups of images [13], human intervention is still needed to connect these processes. The primary objective of this project is to create a tool that can help people automatically cluster images into groups and generate concept-guided descriptions for each collection. This approach would significantly reduce the time and labor required for image data analysis, making it more accessible and efficient for a broader range of applications.

This project combines several computer vision and unsupervised topic-modeling techniques, including vision transformer embeddings, image captioning methods, and neural topic modeling. The project provides an automated pipeline that extracts visual feature embeddings and text-based descriptions from images using vision transformers and image-captioning models. The system uses visual embeddings to generate a pair-wise similarity matrix for visual similarity search. It then combines the image captions and visual embeddings to generate concept-guided topic descriptors using neural topic modeling techniques.

1.2 Review of Literature

The following sections provide an overview of the current research in computer vision related to contextual visual similarity, vision embeddings, and image captioning. More specifically, it summarizes the advent of pre-trained models such as SWIN Transformers [11] and BLIP Image

Captioning [10]. Additionally, these sections will briefly outline the neural topic-modeling technique described in the BERTopic [6] research.

1.2.1 Contextual Visual Similarity

Image similarity is essential in many applications, such as search engines, medical research, and agriculture. However, the definition of image similarity depends on the context relevant to the task. While recommending images based on features such as color or object similarities is helpful in some applications, other systems may find the overall similarities between the environment represented in the images more useful. As demonstrated in [15], we can create contextualized search criteria for image similarity by providing the system with triplets of images (a query image, a positive example, and a negative example) to develop a model for unsupervised attribute discovery. This model then learns the feature weights for each triplet that reduces the distance between the query image and the positive example while increasing the distance between the query and the negative example.

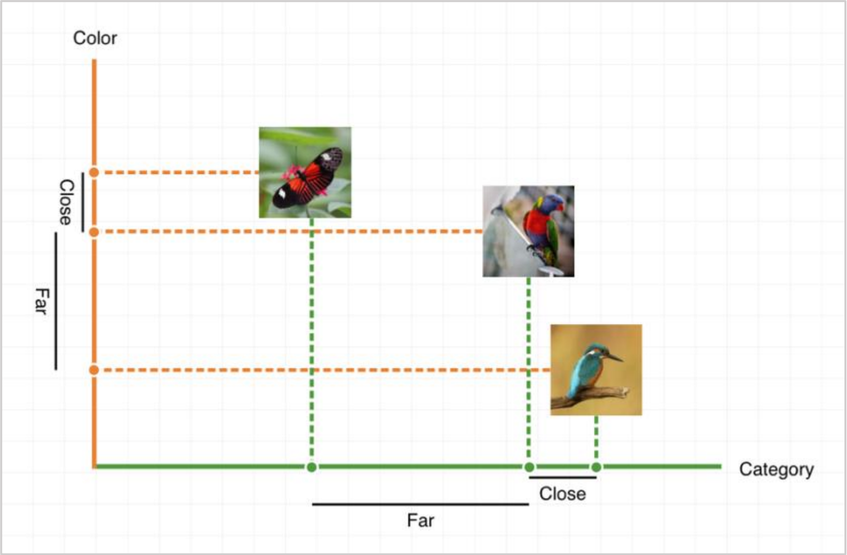

Figure 1: An example of how image similarity depends on context. Adapted from [15] to show how two images of birds are related in terms of their category and unrelated in terms of their color.

The main difference between the methods described in [15] and the methods in this project is that while they require three images to define a contextualized similarity search criterion, ViTopic computes similarity based on the inverse Euclidean distance between vision transformer embeddings. This approach requires only one image embedding per query and does not require the model to learn new feature weights.

1.2.2 Vision Embeddings

Since the inception of the AlexNet model in 2012 [8], computer vision researchers have developed many new models for tasks such as image classification, object detection, and scene segmentation. At the same time, researchers made significant advancements in natural language processing (NLP) by introducing transformers and attention mechanisms [14]. Even though transformers showed promising results in NLP tasks, they faced difficulties in computer vision due to domain differences between the two areas. In 2021, a team of researchers at Microsoft proposed a new vision-transformer model that integrates shifted hierarchical windows of self-attention, which showed improvements over traditional methods [11]. As seen in Figure 2, the localized self-attention mechanism allows these models to pay attention to the important parts of the input signal and learn the abstract feature embeddings that describe the images. Since these models can identify abstract feature maps numerically, we can treat the embeddings as vectors in n-dimensional space and compute their similarity (therefore computing the similarity between the images) using some similarity function.

1.2.3 Image Captioning

Image captioning is a deep learning task that bridges the gap between computer vision and natural language processing by teaching machine learning models to understand the visual content of an image and generate relevant descriptions in natural language. Pre-trained image captioning models make it easier for researchers and developers to build language-vision applications without the need to train their models from scratch. BLIP is an example of a pre-trained model that has shown unparalleled results in multiple metrics, including a 2.7% increase in average recall@1 for image-text retrieval and a 1.6% increase in visual question-answering scores [10]. These metrics indicate the success of the BLIP model in generating more accurate and relevant image captions, making it a valuable tool for researchers and developers looking to build applications that unify language and vision.

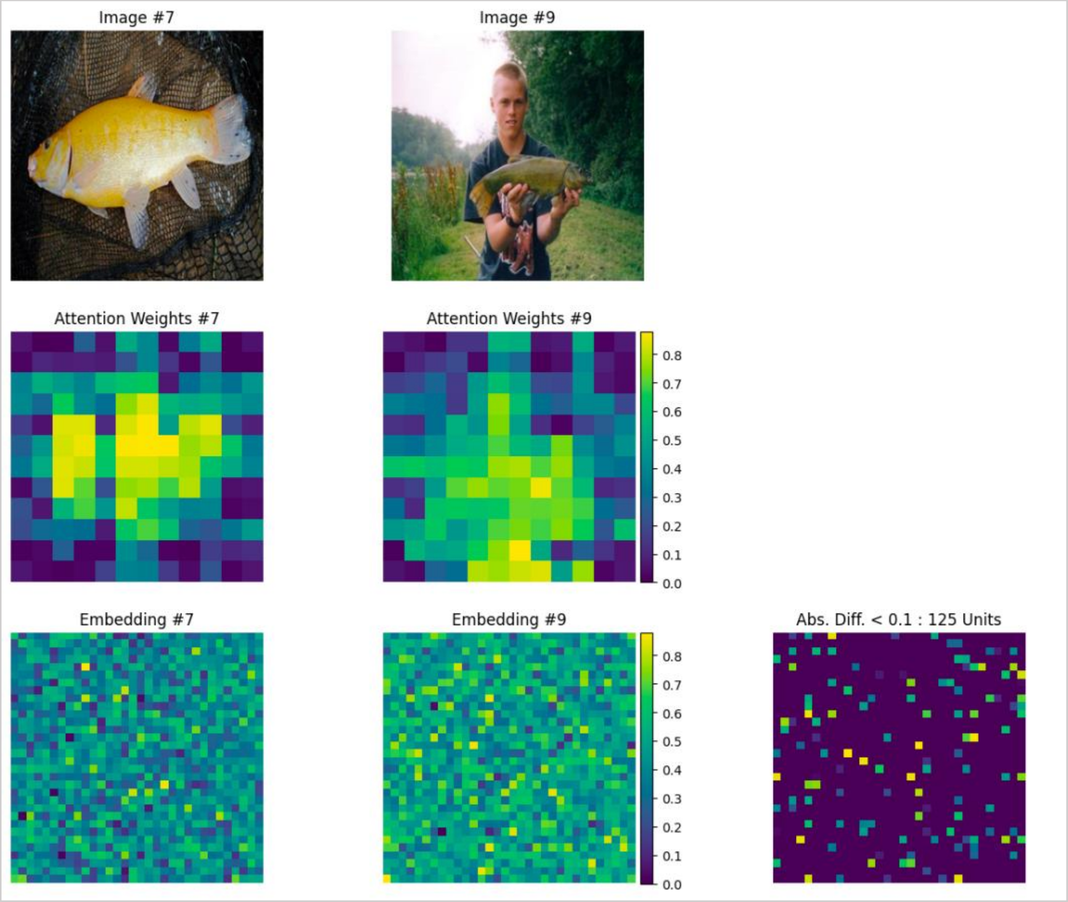

Figure 2: Heatmaps showing areas of highest attention (second row) and output embeddings (third row, first two plots). The last heatmap (third row, third plot) shows the units where the absolute difference between the two output embeddings is less than 0.1 (125 units).

1.2.4 Topic Modeling on Images

Topic modeling is a technique used in natural language processing that aims to identify the hidden structure or topics within a collection of documents. We can extend this idea to images by identifying the collection of images that all share the same *concept*. While traditional topic modeling techniques like Latent Semantic Analysis [9] and Latent Dirichlet Allocation [3] perform well in most contexts, newer techniques such as Neural Topic Modeling with BERTopic [6] allow expanding the idea beyond text. In short, BERTopic works by clustering output embeddings from deep neural networks and generating topic descriptions from a class-based term-frequency inverse-document-frequency matrix. This approach has shown significant results in topic coherence and diversity when compared to traditional methods.

Chapter 2

2 Methods

ViTopic automates the visual modeling pipeline by combining computer vision and natural language processing techniques to extract visual and textual features from a collection of images. It uses vision embeddings to generate a similarity matrix for visual similarity search and combines them with image captions to create a concept-guided topic model. The following sections explain these components and how they contribute to the overall pipeline in more detail.

2.1 Data Processing: Embeddings & Captions

In this project, we used images from the ImageNette dataset, which is a subset of the ImageNet project [5]. The ImageNette dataset consists of 13,394 images across ten categories, providing researchers with a large and diverse enough dataset to train and evaluate their models more quickly. For this project, however, we selected a random sample of 480 images from each of the ten classes, resulting in an even smaller dataset that allowed us to quantify the system's performance using the ground-truth labels present in the dataset.

In the data pre-processing step of the ViTopic pipeline, we resize the images to a consistent size and adjust the color channels if necessary to ensure they are in the correct format. Once we have pre-processed the images, we pass them through a SWIN transformer model [11] to generate image embeddings and capture the high-level visual features of the images. In addition to generating visual embeddings, we use the BLIP image-captioning model [10] to generate textual descriptions for each image. These pre-trained models (microsoft/swin-base-patch4-window12-384-in22k and Salesforce/blip-image-captioning-large, respectively) and the ImageNette dataset can be easily accessed from the HuggingFace repository using their publicly available "Datasets" library in Python.

2.2 Visual Similarity Search

The first part of the project involves developing a system for visual similarity search, which is a critical task in many applications, including content-based image retrieval and image classification. We start by treating the vision embeddings as vectors to capture the high-level visual features of the images and to perform mathematical operations on them.

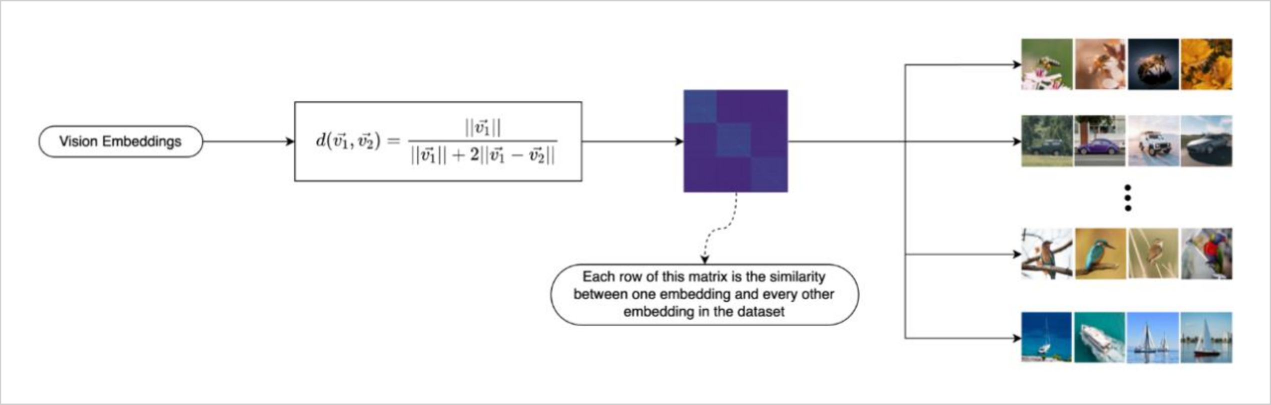
Vision Embeddings
$d(\vec{v_1}, \vec{v_2}) = \frac{||\vec{v_1}||}{||\vec{v_1}|| + 2||\vec{v_1} - \vec{v_2}||}$
Each row of this matrix is the similarity between one embedding and every other embedding in the dataset

Figure 3. A diagram showing the components of the visual similarity search system.

While any similarity function could work for computing the similarity matrix, this project uses the following "query-scaled" version of the inverse Euclidean distance to measure embedding similarity, where x_1 and x_2 are vision embeddings:

$$d(x_1, x_2) = \frac{\|x_1\|/2}{(\|x_1\|/2) + \|x_1 - x_2\|} = \frac{\|x_1\|}{\|x_1\| + 2\|x_1 - x_2\|}$$

By generating a pair-wise similarity matrix using the inverse Euclidean function, the system can compare the visual embedding of the query image to the visual embeddings of all other images in the dataset and rank them based on their score. Users can then select the top K entries relevant to their query.

2.3 Concept-Guided Topic Model

The second part of the project involves generating text-based descriptors for images that share the same concepts. Since the researchers in [1] showed that using UMAP [12] to reduce highdimensional embeddings can enhance the performance of many clustering algorithms, the first step in the concept-guided topic model is to use UMAP to reduce the embeddings from 1024 dimensions to 256. Then, similar to the methods found in [6], we use the HDBSCAN [4] algorithm to assign clusters to each vision embedding, resulting in high-level topics that capture conceptual similarity.

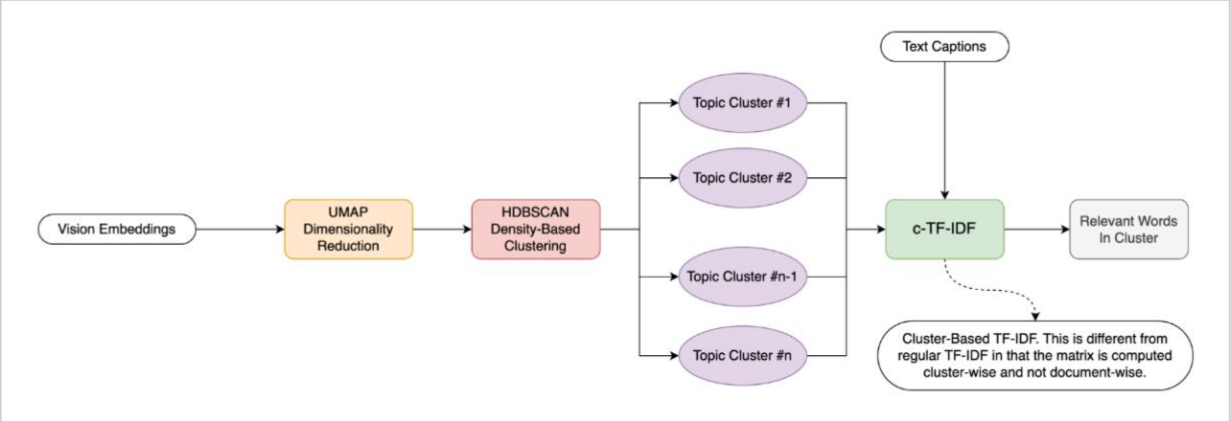
Vision Embeddings
UMAP Dimensionality Reduction
HDBSCAN Density-Based Clustering
Topic Cluster #1
Topic Cluster #2
Topic Cluster #n-1
Topic Cluster #n
Text Captions
c-TF-IDF
Relevant Words In Cluster
Cluster-Based TF-IDF. This is different from regular TF-IDF in that the matrix is computed cluster-wise and not document-wise.

Figure 4. A diagram showing the components of the concept-guided topic model, following a similar approach to the BERTopic [6] architecture.

Finally, using the captions generated in the pre-processing step in conjunction with the cluster information, we generate a class-based term-frequency inverse-document-frequency (c-TF-IDF) matrix that encodes the frequency of words in the clusters relative to other clusters. We can then retrieve the top N descriptors relevant to each cluster of images, effectively describing the overall concept they represent.

Chapter 3

3 Results and Evaluation

In this section, we present the results of our experiments on the visual similarity search and concept-guided topic model components of the project. We evaluate the system's precision and overall accuracy and provide insights into its capabilities and limitations. The metrics we used to measure the performance of ViTopic include Precision@K of the visual similarity search and the collision entropy of the topic clusters using the ImageNette ground-truth labels.

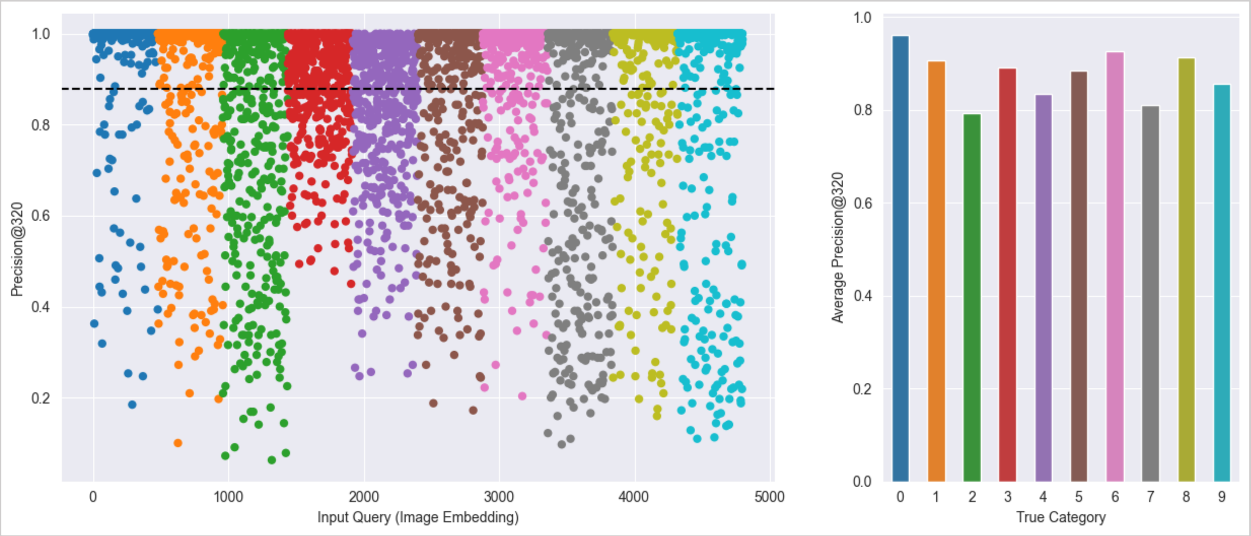

Figure 5: Precision@320 scores for each image in the dataset (left) colored by ground-truth category, and average precision@320 for each ground-truth category (right), reaching an average score of 87.79%.

Since the ImageNette dataset contains ground-truth labels for each of our images, we can quantify the performance of the visual similarity search by treating it as a recommender system and computing its precision at the top K results. Figure 5 shows that the visual similarity search performed well on most images, with an average precision@320 of 87.79% and a variance of 4.07% across all ten categories. We evaluated the model on the top 320 results per query because each category contains 480 images. However, we should note that it is not a perfect solution since many queries scored below 40%. One potential explanation for this is that using the “query-scaled” inverse Euclidean distance assumes that similar vectors live within a perfect hypersphere, which is not at all true for natural images.

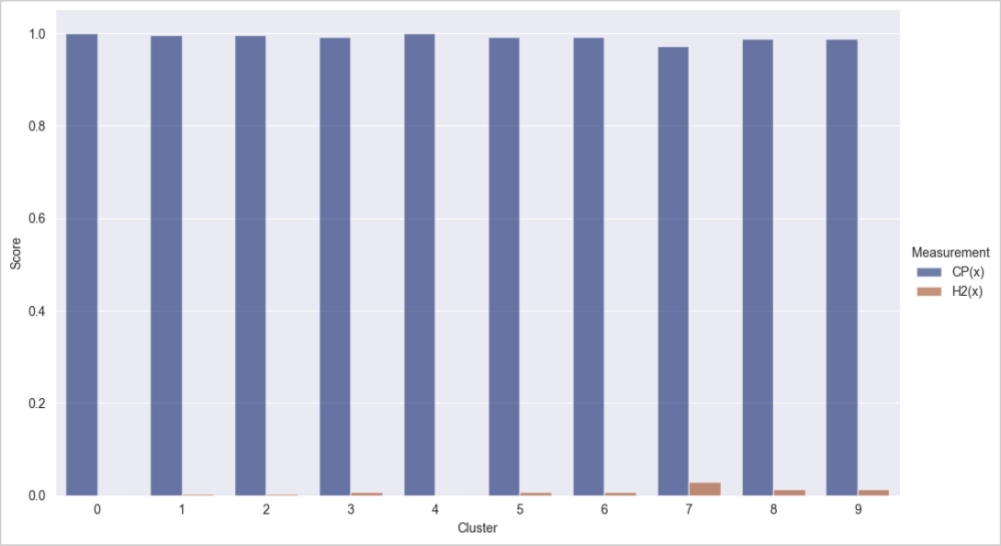

Figure 6: A plot of the collision probability $CP(x)$ and collision entropy $H_2(x)$ for the clusters discovered by the concept-guided topic model.

On the other hand, the concept-guided topic model showed outstanding performance, accurately generating the expected number of clusters and topic descriptors in line with the ImageNette dataset. To evaluate the homogeneity of the clusters based on the ground-truth ImageNette labels, we employed two measures: collision entropy $H_2(x) = -\log \sum_{i=1}^{n} p_i(x)^2$, and collision probability $CP(x) = \sum_{i=1}^{n} p_i(x)^2$. A high CP score indicates that most of the elements in a cluster belong to the same category, while a low H_2 score denotes clusters with low diversity and high homogeneity. Specifically, we observed that the concept-guided topic model produced high CP scores (99.13% on average) and low H_2 scores (0.87 on average), signifying that it effectively grouped images with similar concepts together. The following list shows the top five descriptors for each of the clusters in order of decreasing relevance:

- **Cluster 0**: golf, ball, close, balls, grass.
- **Cluster 1**: fish, holding, man, caught, hands.
- **Cluster 2**: dog, laying, grass, standing, sitting.
- **Cluster 3**: truck, araf, parked, garbage, road.

- **Cluster 4**: radio, close, player, table, stereo.
- **Cluster 5**: parachute, flying, sky, araffe, air.
- **Cluster 6**: church, steeple, clock, tower, view.
- **Cluster 7**: horn, french, brass, playing, instrument.
- **Cluster 8**: chainsaw, tree, cutting, man, handle.
- **Cluster 9**: gas, pump, station, pumps, sitting.

By qualitatively evaluating the topic descriptors, we observe that the results of the concept-guided topic model are promising since it successfully identified coherent clusters of images with high homogeneity. The system provided meaningful topic descriptors that aligned well with the ImageNette dataset, suggesting that the concept-guided approach may be helpful for image clustering and topic modeling in other contexts, particularly when the underlying categories or concepts are well-defined.

3.1 Cluster-Weighted Visual Similarity Search

One of the problems with the visual similarity search is that using the “query-scaled” inverse Euclidean distance assumes that the embeddings form perfect hyperspheres. However, this idea is not accurate for natural images, especially in the case of a single embedding, which the function would assume is always at the center of the hypersphere. Consider the specific example of an image of a man holding a fish (x_1). This image contains two objects: the man and the fish. Now consider another image with a man holding a chainsaw (x_2). Both images have a common subject - a man - which would place them close to each other in the embedding space. However, they also belong to two categories or topics - fishing and chainsaws - meaning that the images lie on the boundary between the two topics in the embedding space. If we assume that x_1 is the query, then using the “query-scaled” inverse Euclidean distance would result in retrieving images from both topics, even though the ground-truth labels assign them to different categories.

The challenge is ensuring the recommended items are visually similar *and* conceptually relevant to the query. One solution to address this problem is to incorporate the cluster information obtained from the concept-guided topic model. We can effectively filter out items that are visually similar but irrelevant to the query by penalizing the similarity scores of out-of-cluster items.

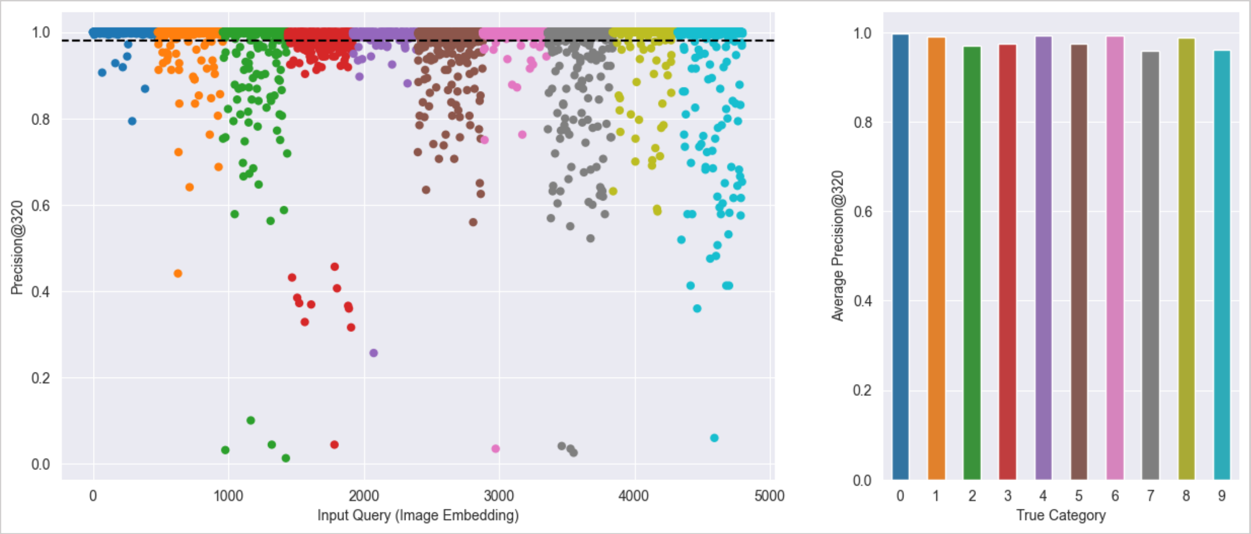

Figure 7: Precision@320 scores for each image in the dataset (left) colored by ground-truth category, and average precision@320 for each ground-truth category (right), reaching an average score of 98.01%.

Our experiments showed that penalizing the out-of-cluster scores by only 10% significantly improved the visual similarity search. Specifically, the average score increased by 11.64%, reaching an average precision@320 of 98.01% with a variance of 0.62%, as shown in Figure 7. The results demonstrate the effectiveness of the cluster information in guiding the search toward conceptually relevant items.

It is worth noting that increasing the penalty further could potentially boost the scores to 100%, indicating that all recommended items are within the same cluster as the query. However, this approach would not be practical in real-world scenarios where users may be interested in exploring visually similar but conceptually diverse items. Therefore, a balanced penalty that filters out irrelevant items while allowing for some diversity in the recommended items would be more desirable.

4 Conclusion

In this project, we developed ViTopic: a unified system for topic modeling on images by clustering vision transformer embeddings. This report includes a brief overview of the main components of ViTopic – visual similarity search and concept-guided topic modeling – and outlines the results of evaluating the components using metrics such as precision of top K and collision entropy. Overall, the results of this project suggest that ViTopic has the potential to be a valuable tool for a wide range of image-related tasks, from content management and organization to automated image tagging and recommendation systems. Future work could explore other ways to improve the system's accuracy and efficiency.

References

[1] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In *Image and Signal Processing* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 317–325. DOI:https://doi.org/10.1007/978-3-030-51935-3_34

[2] Souad Azzouzi, Amal Hjouji, Jaouad EL-Mekkaoui, and Ahmed El Khalfi. 2023. A novel efficient clustering algorithm based on possibilistic approach and kernel technique for image clustering problems. *Appl. Intell.* 53, 4 (February 2023), 4327–4349. DOI:https://doi.org/10.1007/s10489-022-03703-0

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.

[4] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 160–172. DOI:https://doi.org/10.1007/978-3-642-37456-2_14

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. DOI:https://doi.org/10.1109/CVPR.2009.5206848

[6] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (2022). DOI:https://doi.org/10.48550/ARXIV.2203.05794

[7] Oded Hupert, Idan Schwartz, and Lior Wolf. 2022. Describing Sets of Images with Textual-PCA. (2022). DOI:https://doi.org/10.48550/ARXIV.2210.12112

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. DOI:https://doi.org/10.1145/3065386

[9] Tom Landauer and Scott Dooley. 2002. Latent semantic analysis: theory, method and application. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community* (CSCL' 02), International Society of the Learning Sciences, Boulder, Colorado, 742–743.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Retrieved April 30, 2023 from http://arxiv.org/abs/2201.12086

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Retrieved April 30, 2023 from http://arxiv.org/abs/2103.14030

[12] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. DOI:https://doi.org/10.48550/arXiv.1802.03426

[13] Andrea Pasini, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2022. Semantic Image Collection Summarization With Frequent Subgraph Mining. *IEEE Access* 10, (2022), 131747–131764. DOI:https://doi.org/10.1109/ACCESS.2022.3229654

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (2017). DOI:https://doi.org/10.48550/ARXIV.1706.03762

[15] Xiaofang Wang, Kris M. Kitani, and Martial Hebert. 2016. Contextual Visual Similarity. Retrieved April 30, 2023 from http://arxiv.org/abs/1612.02534

[16] Feng Zhang, Lin Li, Qiang Hua, Chun-Ru Dong, and Boon-Han Lim. 2022. Improved deep clustering model based on semantic consistency for image clustering. *Knowl.-Based Syst.* 253, (October 2022), 109507. DOI:https://doi.org/10.1016/j.knosys.2022.109507